Comparison of Sequence Variation of H5N1 Influenza a Virus using Wavelets

Shiwani Saini
Department of Electrical Engineering
National Institute of Technology
Kurukshetra, India

Lillie Dewan
Department of Electrical Engineering
National Institute of Technology
Kurukshetra, India

ABSTRACT

With the growing volume of deoxyribonucleic acid (DNA) sequence database of the human and model organisms since the completion of Human Genome Program (HGP), characterizing the information contained in such sequences is of utmost importance. In this work, wavelet transforms have been used to determine the variations in the genomic sequences of Influenza A virus. Nucleotide sequences of all the 8 segments of H5N1 virus occurring in different regions, in different hosts and over different years have been compared to determine the sequence variations.

General Terms

Wavelets, Signal Analysis, Influenza A

Keywords

Influenza A virus, wavelet transforms, genomic sequences, sequence variability, Electron ion interaction pseudo potentials.

1. INTRODUCTION

Influenza A viruses belong to the Orthomyxoviridae family. A typical feature of influenza A viruses is that its genome is divided in eight distinct linear segments of negative-sense single stranded RNA [1] including: HA (hemagglutinin), NA (neuraminidase), NP (nucleoprotein), M (two matrix proteins, MI and M2), NS (two distinct non-structural proteins, NS1 and NEP), PA (RNA polymerase), PB1 (RNA polymerase and PB1-F2 protein), and PB2 (RNA polymerase) [2]. Of all influenza proteins, mutations in hemagglutinin (HA) and neuraminidase (NA) show significant variations in their sequences [3]. The segmented nature of the genome favours the exchange of entire genes between different viral strains. This increases the probability of mutations between strains cohabitating the same cell.

In general, an influenza virus infects only a single species; however, whole viruses may occasionally be transmitted from one species to another, and genetic reassortment between viruses from two different hosts can produce a new virus capable of infecting a third host [4] and thus causing a pandemic. Influenza pandemic is continuing to rise due to its constant antigenic drift. The severity of the disease and the number of deaths caused by a pandemic virus varies greatly and can change over time. These changes can be interpreted by graphical methods using signal processing methods.

Genomic information can be converted into digital form by representation of the nucleotide bases in the form ofmathematical sequences [5]. The use of DSPprinciples to analyze genomic sequences requires defining an adequate representation of the nucleotide bases by numerical values, converting the nucleotide sequences into an equivalent of time series [6] wherein nucleotide bases are represented on the x-axis and y-axis represents the mathematical values assigned to the sequence.

There are several mathematical transforms such as Fourier Transforms, Short Time Fourier Transforms, Wavelet Transforms, Hilbert Transforms, etc. which can be applied to the digital data. These mathematical transformations when applied to the digital signals are used to obtain information from that signal that is not readily available in the raw signal [7]. Transformed signals can be used for graphical representation of genomic data and offers the advantage of analyzing the characteristics and regions of interest within the DNA sequence without the use of statistical techniques and pattern matching methods. Several methods can be used for determining the variations of sequences that include molecular characterisation [8, 9], Z curve method [10], phase analysis of genomic sequences [11], graphical sliding window techniques [12].

In the present work, the variations undergone by viral strains on account of mutations of various proteins of Influenza A virus have been determined. The nucleotide sequences of different viral strains have been converted into mathematical representations and transformed using discrete wavelet transform. Visual comparison of the plots of wavelet coefficients of different sequences show significant variations in sequences occurring in different regions and in different hosts.

2. WAVELET TRANSFORMS

Wavelet is a waveform of finite duration and zero average value. Wavelet transform (WT) is obtained using a wavelet function $\psi(t)$, in which the original signal is convolved with the scaled and shifted version of the mother wavelet. Wavelet transforms are capable of transforming the signal simultaneously in both time and frequency domain hence offer the advantage of time frequency localization of any event. Wavelet analysis can be applied to one dimensional data, two-dimensional data (images) and, in principle, to higher dimensional data.

There are two types of wavelet transforms:
1. Continuous wavelet transforms (CWT)
2. Discrete wavelet transforms (DWT)

2.1) Continuous Wavelet Transforms

The CWT compares the signal to shifted and compressed or stretched versions of a wavelet. Stretching or compressing a function is collectively referred to as dilation or scaling. By comparing the signal to the wavelet at various scales and positions, function of two variables (scale and position) is obtained. For a scale parameter, a>0, and position, b, the CWT is:

$$Cab = \int_t f(t) \frac{1}{\sqrt{a}} \frac{\psi(t-b)}{a} dt$$

By continuously varying the values of the scale parameter, a, and the position parameter, b, CWT coefficients C(a,b) are obtained.

2.2) Discrete Wavelet Transforms

Continuous wavelet transforms generate a large amount of data as the transform is calculated at all possible scales and positions. In discrete wavelet analysis, scales and positions are chosen based on powers of two called the dyadic scales. After discretization the wavelet function is defined as:

$$\psi_{m,n}(t) = \frac{1}{\sqrt{a_0^m}} \psi * \left(\frac{t - nb_0 a_0^m}{a_0^m}\right)$$

where a0 and b0 are constants. The scaling term is represented as a power of a0 and the translation term is a factor of a0^m. The most common choice for the parameters a0 and b0 are 2 and 1 (dyadic grid scaling). The dyadic grid wavelet is expressed as:

$$\psi_{m,n}(t) = \frac{1}{\sqrt{2^m}} \psi\left(\frac{t - n2^m}{2^m}\right) = 2^{-m/2} \psi(2^{-m} t - n)$$

An efficient way to implement this scheme using filters was developed by Mallat [13] in 1988. The most basic filtering process is represented by (Figure 1).

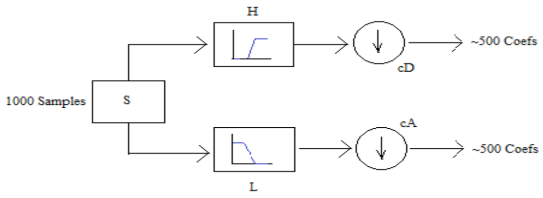

Fig 1: Signal decomposition using DWT

The original signal passes through a pair of high pass and low pass filters, is down sampled to get the decomposed signal through each filter which is half the length of the original signal. The signal S can be expressed as S = cD + cA.After the analysis of the signal, the original signal can be synthesised using inverse discrete wavelet transform. The signal is reconstructed as shown in Figure 2. Reconstruction involves the up sampling of the decomposed signal followed by filtering through two complementary filters.

The complete decomposition and reconstruction process of the signal with the low- and high-pass decomposition filters (L and H) and their associated reconstruction filters (L' and H'), form a system of quadrature mirror filters (Figure 3).

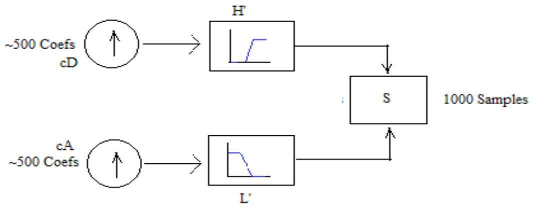

Fig 2: Signal reconstruction using DWT

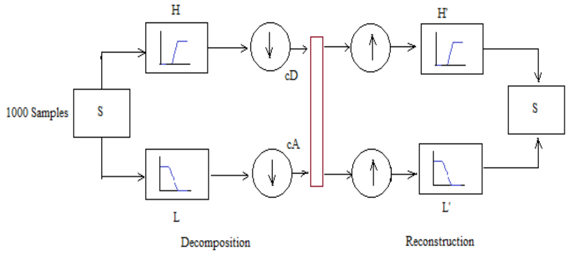

Fig 3: Signal decomposition and reconstruction

3. GENOMIC SIGNAL REPRESENTATION

The main nucleic genetic material of cells is represented by DNA molecules that have double helix structure comprising of two antiparallel intertwined complementary strands, each a helicoidally coiled heteropolymer. Four kinds of nitrogenous bases are found in DNA that constitute the genomic sequences: thymine (T) and cytosine (C)—which are pyrimidines, adenine (A) and guanine (G)—which are purines. A pyrimidine in one chain always faces a purine in the other along the two strands of DNA double helix, and only the base pairs T−A and C−G exist. As a consequence, the two strands of a DNA helix are complementary, store the same information, and contain exactly the same number of A and T bases and the same number of C and G bases.To express the genomic sequences mathematically, there are several methods such as Voss representation [14], purine (A, G = -1) and pyrimidines (C, T = +1) representation [15], mapping of the nucleotides onto a complex tetrahedral plane [16], complex number representation [17], electron ion interaction potential (EIIP) [18] and integer number representation (Bioinformatics toolbox in Matlab).

4. METHOD

Different viral nucleotide sequences of H5N1 taken from the different hosts occurring in different years and in different regions have been downloaded from NCBI database [19]. The sequences were represented in mathematical form using electron ion interaction pseudo potentials (EIIP) and were then transformed by discrete wavelet transform using Daubechies’ wavelet (dB4). The decomposition level was chosen to be 5. Comparison of the transformed sequences at different levels can be used to investigate the variation in the sequence composition of different strains of Influenza virus. The original sequences were transformed using DWT. The multi resolution analysis upto level 5 decomposes the sequences into approximations and details at various levels (1-5). Whereas wavelet coefficients at coarse scales correspond to low frequency components in the signal (approximations) and capture gross and global features of the signal, wavelet coefficients at fine scales (details) correspond to high frequency information and contain local details. Since approximation coefficients give the global variation of the signals, sequence comparison of various proteins of the virus can be done by plotting the approximation coefficients. These plots identify the trend of the signal. Overlapping curves of approximation coefficients determine the global similarity. The local variation of the signal can be determined by comparing the energies of the wavelet coefficients at various levels. Similarity in the energy plots of detail coefficients corresponds to local similarity in the sequences.

5. RESULTS

The energy plots of detail coefficients at all levels and approximation coefficients plots at level 5 of all the 8 segments of different viral sequences of H5N1 were compared to determinelocal and global trend in sequence variations respectively.

The HA protein sequence plot (Figure 4) shows that sequence AB212054 is significantly different from the rest of the sequences. These differences appear due to the mutations undergone by the sequences across different years of characterization.AB212054 was sequenced in 2003, hence appears different whereas the rest of the sequences were characterized in the same year (1997) and form a cluster with almost similar global trend, showing only slight variations in the peaks along the sequence length. The HA sequence plot in Figure 5 shows that the sequences AF082035, 37 have almost similar plots and differ from the cluster of plots of sequences AF102680,81,82. Even though all these sequences were characterized in the same year (1997), the two different clusters appear because of different hosts. The sequences AF102680, 81, 82 were isolated from humans and the sequences AF082035, 37 were isolated from chickens.

The plot for NA protein (Figure 6) also shows that sequence AB212056 varies from the plots of sequences AF046081, 89. The difference is on account of different years of characterisation of the sequences even though the sequences have been isolated from chickens and occur in the same region (Hong Kong). Thoughthe sequences AF046081, 89 are globally similar, they tend to show local variations in the plots around the region between 400-600 bases that suggest the regions of mutations.

Similar trends in clusters are also evident for NP, PB1, PB2 proteins wherein the plots for sequences in the year 2003 are different from the plots of the sequences isolated in year 1997 occuring in the same hosts and in same regions. The plots of all the M proteins ofInfluenza virus appear different due to the different years of molecular characterization. Coefficients plot of PA sequence (Figure 8) shows that the peaks of all the sequences have shifted suggesting a drift in all the sequences over the years. NS1 sequence plots for AB212058, AF046083, 91 are overlapping suggesting that these have not undergone significant variations. However AB212059 is quite deviatedfrom the rest of the sequences probably due to the different year of sequencing.The local variations in the sequences can also be identified by comparing the energy of the detail coefficients at different levels. Local similarity suggests identical detail coefficients and hence similar energy of the coefficients. Energy of the detail coefficients for all the sequences of HA protein appear different suggesting dismilarity in the entire sequence. Energy of detail coefficients of sequences 2 and 3 of NA, NP, PB1 proteins are similar showing local similarity in the sequences. The energy of all the sequences of PA, M proteins are dissimilar thereby suggesting dissimilarity. PB2 protein shows similar energy in coefficients of sequences 1, 2 possibly due to same year of occurrence but sequences 3 and 4 appear dissimilar, despite the same year of occurrence due to mutations.

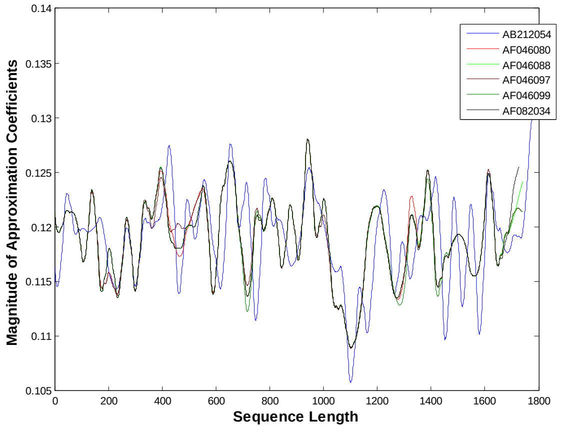

Fig4: Approximation coefficients plot for HA protein

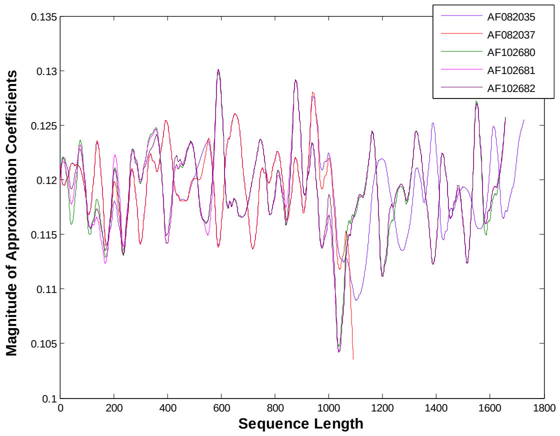

Fig 5: Approximation coefficients plot for HA protein

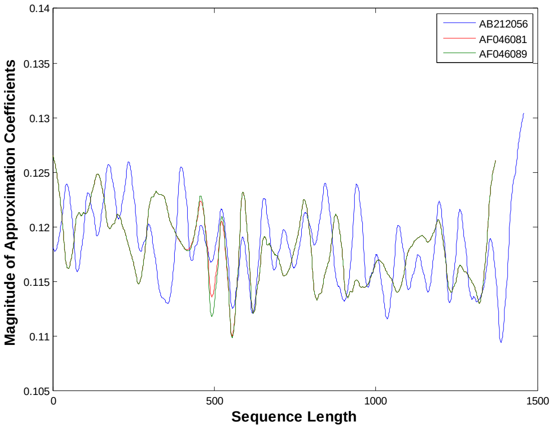

Fig 6:Approximation coefficients plot for NAprotein

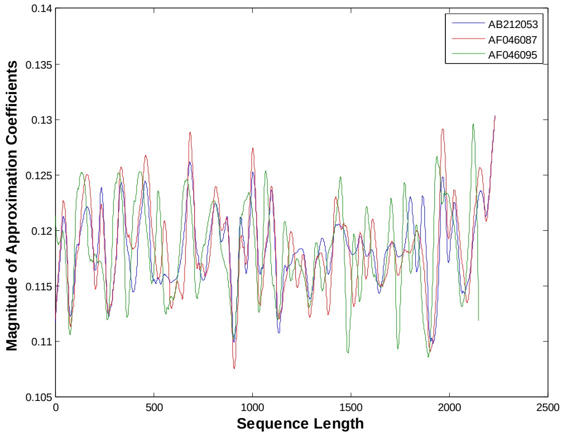

Fig7:Approximation coefficients plot for NPprotein

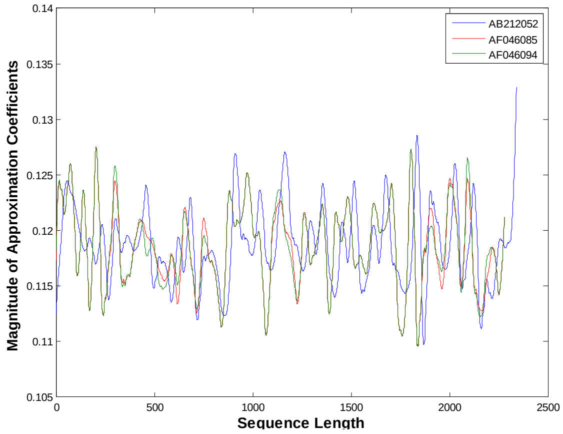

Fig 10: Approximation coefficients plot for PB1 protein

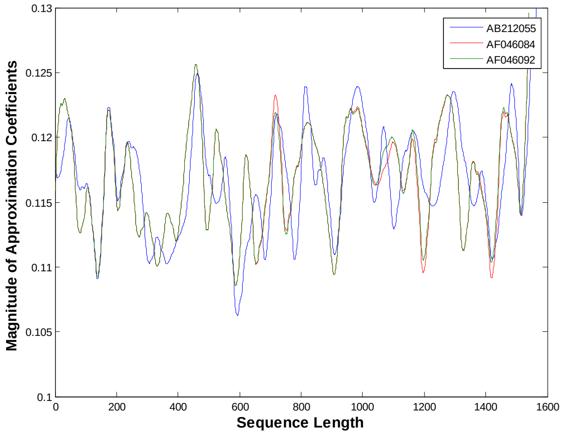

Fig 8:Approximation coefficients plot for PAprotein

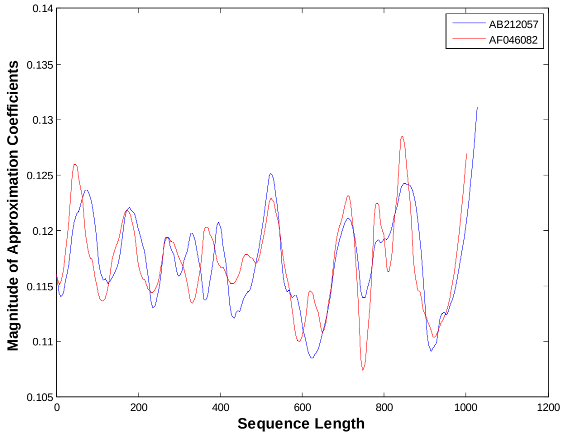

Fig 11:Approximation coefficients plot for PB2protein

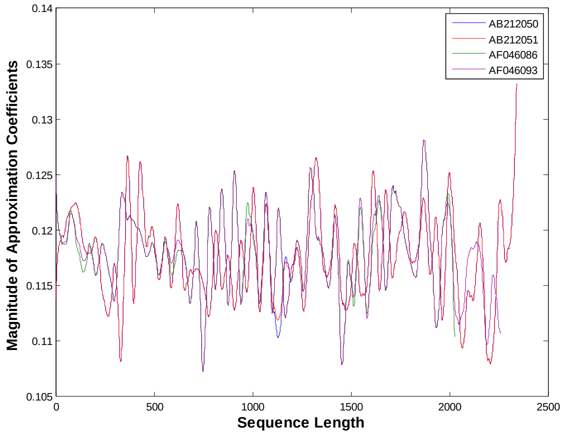

Fig 9:Approximation coefficients plot for Mprotein

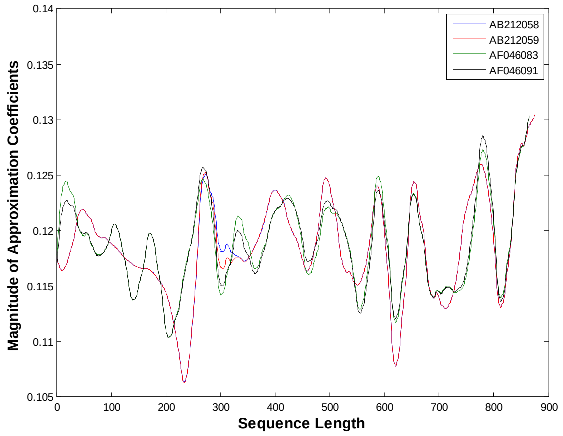

Fig 12:Approximation coefficients plot for NS1protein

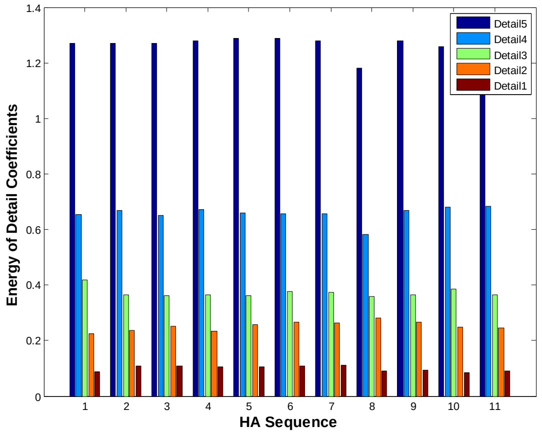

Fig.13 Energy of detail coefficients of HA protein

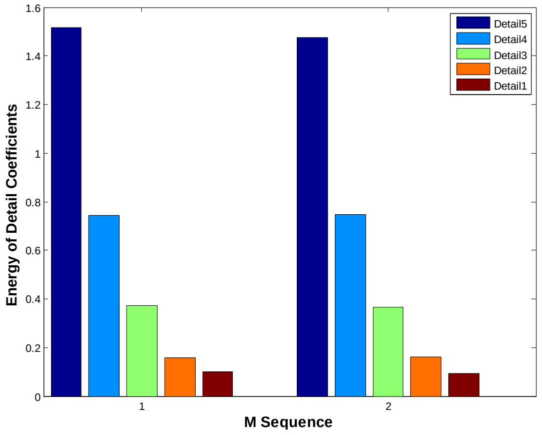

Fig.16 Energy of detail coefficients of M protein

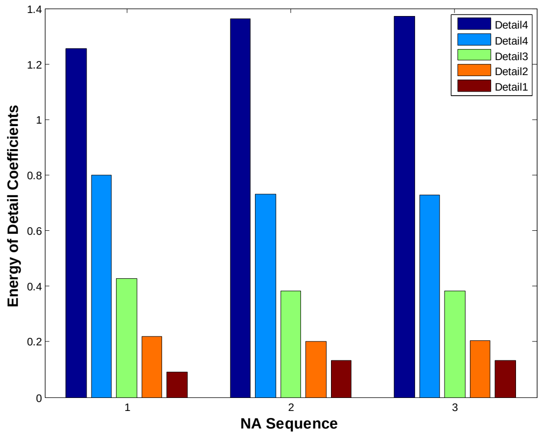

Fig.14 Energy of detail coefficients of NA protein

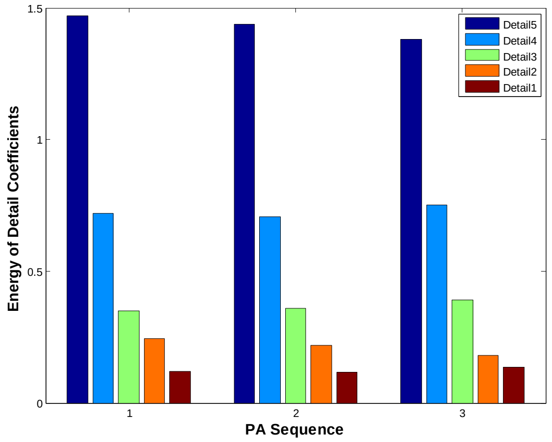

Fig.17 Energy of detail coefficients of PA protein

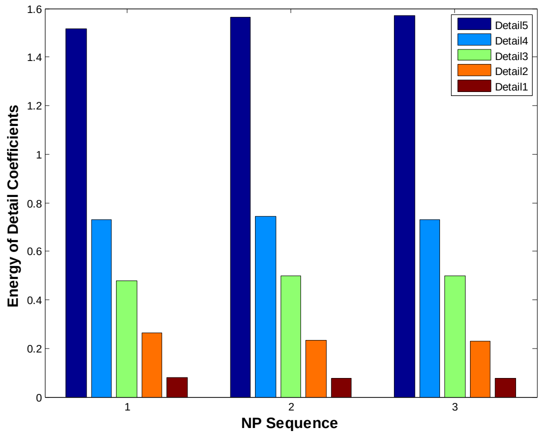

Fig.15 Energy of detail coefficients of NP protein

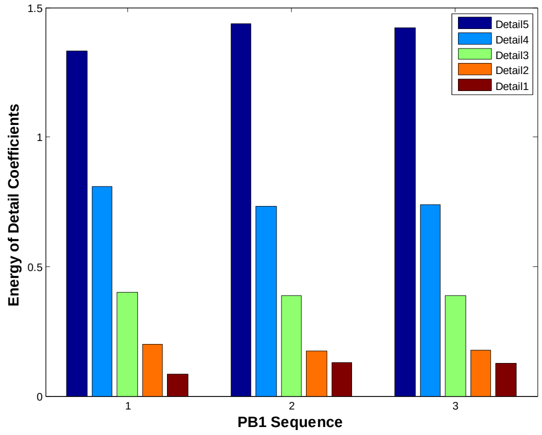

Fig.18 Energy of detail coefficients of PB1 protein

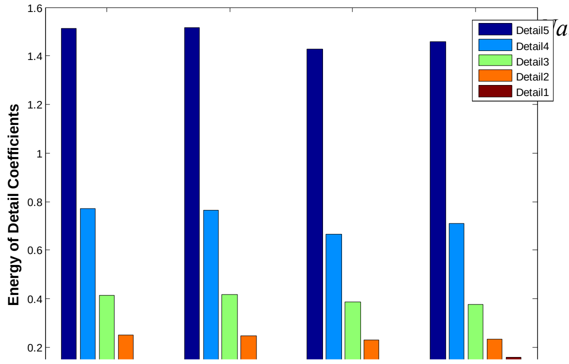

Fig.19 Energy of detail coefficients of PB2 protein

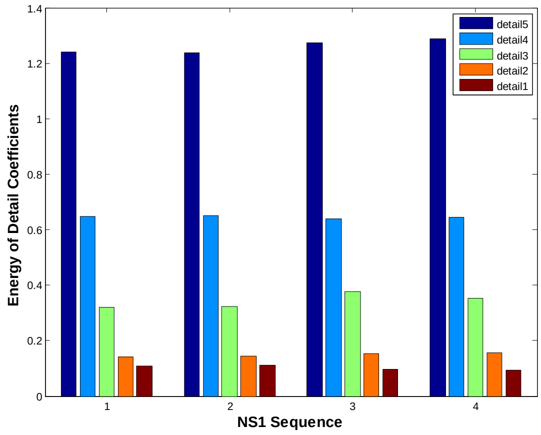

Fig.20 Energy of detail coefficients of NS1 protein

5. CONCLUSION

Wavelet transforms offer the advantage of reducing the computational complexity as compared to mathematical calculations and faster identification of sequence changes by visual representation of the plot of the wavelet coefficients. These positions where the coefficient plots appear different are the regions where the sequences have undergone changes in the nucleotide composition. Relatively stable regions in different protein sequences of H5N1 can be used as the target regions in diagnosis and vaccine manufacturing.Thus the combined use of the digital signal processing methods and bioinformatics provides a simple tool for analyzing the interactions in the viral sequences accumulating in unprecedented large numbers from throughout the world during the epidemics and can be used for vaccine design and new diagnosis development.

6. REFERENCES

[1] Fodor, E., Brownlee, G.G., Potter, C.W., "Perspectives in medical virology, Influenza virus replication", Elsevier (Amsterdam), 2002,1–29.

[2] Claas, E.C., Osterhaus, A.D., and Van Beek, R., "Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus" , Lancet, 351,(1998) 472–477.

[3] Scholtissek, C., Burger, H. ,Kistner, O. andShortridge, K. F. , "The nucleoprotein as a possible major factor in determining host specificity of influenza H3N2 viruses", Virology,147, (1985), 287–294,.

[4] Shinde, V., Bridges, C.B., Uyeki,T.M., Shu, B., Balish, A., Xu, X., et al., "Triple-reassortant swine influenza A (HI) in humans in the United States, 2005-2009", N Engl. J. Med., 360, (2009), 2616-25.

[5] Cristea,P., "Conversion of Nitrogenous Base Sequences into Genomic Signals", Journal of Cellular and Molecular Medicine, 6, 2, (April – June 2002), 279-303.

[6] Juan, V., et al. "Digital Signal Processing in the Analysis of Genomic Sequences", Current Bioinformatics, (2009), 4,28-40.

[7] RobiPolikar. 1999. The Wavelet Tutorial, http://users.rowan.edu/~polikar/WAVELETS/WTtutorial .html.

[8] David, L., Suarez, et al., "Comparisons of Highly Virulent H5N1 Influenza A Viruses Isolated from Humans and Chickens from Hong Kong", Journal Of Virology, (Aug. 1998),6678–6688.

[9] Ghedin ,E.,et al.,"Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution", 437|20 (October 2005)|doi:10.1038/nature04239.

[10] Yan-ling Yang,2009, 229-231. A Geometrical Analysis of the Avian Influenza Virus at different areas. In Proceedings of International Conference on Engineering Computation.

[11] Cristea,P.D., Tuduce,R., 2006.Genomic Signal Analysis of Avian Influenza Virus Variability. InProceedings of Second International Symposium on Communications, Control and Signal Processing.

[12] Ghosh et al, "Computational analysis and determination of ahighly conserved surface exposed segment inH5N1 avian flu and H1N1 swine flu neuraminidase", BMC Structural Biology,10: 6 (2010).

[13] MallatS., 2000.A Wavelet Tour of Signal Processing, Second edition, Academic Press, New York.

[14] VossR. F., "Evolution of Long-range Fractal Correlations and 1/f noise in DNA base sequences," Physical Review Letters, 68 (June 1992), 3805-3808.

[15] SwarnaBaiArniker and Hon Keung Kwan. Graphical Representation of DNA sequences. In Proceedings of IEEE International Conference on Electro/Information Technology, EIT 2009, 311-314.

[16] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals," J. Cell. Mol. Med., vol. 6, no. 2, pp. 279–303, 2002.

[17] Berger,J.A.,Mitra, S.K., Carli,M.,Neri,A. (11–13 October 2002).New approaches to genome sequence analysis based on digital signal processing. In Workshop on Genomic Signal Processing and Statistics (GENSIPS).

[18] Nair, A.S., and Sreenadhan, S.P., "A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)", Bioinformation 1(6),(2006), 197-202.

[19] The National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine website http://www.ncbi.nlm.nih.gov/genoms/,ftp://ftp.ncbi.nlm. nih.gov/genoms/,GenBank, http://www.ncbi.nlm.nih.gov/Genbank/index.html.